DOI: 10.14421/ijid.2020.09201 IJID (International Journal on Informatics for Development), *e-ISSN: 2549-7448*
Vol. 9, No. 2, 2020, Pp. 59-65

Implementation of Web Scraping to Build a Web-Based Instagram Account Data Downloader Application

Arif Himawan[1], Adri Priadana[2], Aris Wahyu Murdiyanto[3]
Department of Information Systems
Faculty of Engineering and Information Technology
Universitas Jenderal Achmad Yani Yogyakarta
Yogyakarta, Indonesia
[1]reef1881@gmail.com, [2]adripriadana3202@gmail.com, [3]ariswahyumurdiyanto@gmail.com

Article History
Received Sept 10th, 2020
Revised Sept 27th, 2020
Accepted Sept 27th, 2020
Published Oct, 2020

***Abstract*— Instagram has been used by many groups, such as business people, academics, to politicians, to take advantage of the insights gained by processing and analyzing Instagram data for various purposes. However, before processing and analyzing data, users must first pass data collection or downloading from Instagram. The problem faced is that most data collection methods are still done manually. On the other side, many parties offer Instagram account data download service, but it is not entirely free. This research applied a web scraping method to automatically build a web-based Instagram account data download application so that several parties can use it. The web scraping method was chosen because by using this method, researchers do not need to use Instagram's Application Programming Interface (API), which has access restrictions in retrieving data on Instagram. In this study, application testing was conducted on 15 Instagram accounts with various publications between 100 and 11000. Based on the download data analysis results, the web scraping method's application to download Instagram account data can successfully download a maximum of 2412 account data. In this application, users can download Instagram account data to Data Collection and then manage it like deleting and exporting data collection in CSV, Excel, or JSON.**

Keywords*— *data download application; Instagram account data; web scraping; social media; Instagram

1 INTRODUCTION

Currently, Instagram has become an essential part of everyday life for people in the world. Instagram has become one of the fastest-growing social media platforms in recent years [1]. Based on data obtained from a Techcrunch site, Instagram has become one of the fastest-growing social media platforms, namely 1 billion users in June 2018 [2]. It makes Instagram used by many parties from various fields.

Instagram has become a part of the life of today's technology-conscious society. Instagram has been used in various life domains, from economy/business, education, politics, and so on. It has made many groups, ranging from business people, academics, to politicians, to take advantage of the insights obtained by processing and analyzing Instagram data for various purposes such as data mining needs. Data mining is an important method to increase efficiency in finding new or hidden information useful, valid, and easy to understand from extensive databases [3]. It can be more practical, such as finding hidden patterns or rules within the scope of these data sets [4]. However, before processing and analyzing data, users must first retrieve or download data from Instagram. The problem is that most data collection methods are still done manually. Many parties offer Instagram account data download services, but it is not entirely free such as ScrapyGram [1], Octoparse[2], ScrapeStorm[3], and PhantomBuster[4].

This study aims to implement the web scraping method in building a web-based Instagram account data downloader application. The web scraping method automatically takes Instagram account data so that several parties can use it without paying (free). The web scraping method was chosen because by using this method, researchers do not need to use Instagram's Application Programming Interface (API), which has access restrictions in retrieving data on Instagram.

Research on downloading Instagram account data has been conducted before by Sa'dyah, et al., in 2019 [5]. In their study, a crawler engine was built to get Instagram data by utilizing Apache Spark and Apache Kafka. The required data was retrieved through the Application Programming Interface (API) on Instagram. Apache Kafka handles data streaming management, while Apache Spark is used to process data selection and visualization. Yadranjiaghdam et al., in 2017 [6], also implemented Apache Spark and Apache Kafka to develop a real-time Twitter data analysis framework. In their research, Apache Kafka is used to stream data from the Twitter API, while Apache Spark is used for real-time data processing. In this research, the researcher applied the web scraping technique, which was chosen because researchers did not need to use Instagram's API, which has access restrictions in retrieving data on Instagram.

Previous researchers have never made a specific and independent application to download Instagram account data. In general, the data download process is carried out simultaneously at the time of the research. These studies utilize Instagram account data for various purposes, such as determining the effectiveness of promotions, measuring engagement rates, and so on. Fauziah et al., in 2018 [7], used post data from an Instagram account to measure the effectiveness of promotions from the Mount Pancar recreational tourist destination. Promotion through Instagram tends to be considered useful as a new medium for marketing promotions because it has many users. Another study conducted by Arman & Sidik in 2019 [8] used data from several Ministry Instagram accounts and Indonesian Government Institutions to measure the engagement rate where data download was carried out using the web scraping technique. The correlation analysis results from this study indicate that the more the number of followers, the lower the level of engagement. Therefore, the grouping is a range of engagement rate values based on the number of followers.

Another study conducted by Akrianto et al. in 2019 [9] used Instagram account data to determine the best endorsement account. According to him, Instagram account data such as the number of followers, the number of likes, the number of comments, and the level of update updates are the best parameters for choosing the best endorsement account on Instagram. Utama and Inayati, in 2019 [10], used Instagram account data to analyze and categorize brand posts from the official Instagram accounts of the three largest car market share industries in Indonesia. According to him, a company should encourage creative content such as videos on Instagram accounts to attract more viewers and increase engagement rates.

Kurniawan et al., in 2019 [11], studied how to classify Instagram's hashtags using intersection operators and apply the intersection operator. Azmi et al., in 2020 [12], used Instagram posts data related to tourist destinations to find out the reputation of tourist destinations on the island of Lombok. Priadana and Murdiyanto in 2020 [13] used the web scraping method to extract data from Instagram accounts and then process it to analyze the best time. This research has succeeded in using the web scraping method to extract data from 10 trip provider accounts on the Instagram platform.

2 METHOD

In this study, Instagram account data was extracted using the web scraping method. This study's stages consisted of analyzing Instagram account data, implementing the web scraping method, testing data download applications, and analyzing application user interactions.

2.1 Instagram Account Data Analysis

Instagram is a popular social media [14] that is used to share images [15]. Instagram allows users to upload and share pictures and videos with their followers or a selected group of friends [16]. Instagram allows interaction between users, such as through following, liking, and writing comments. Each account

[1] https://www.scrapygram.co/
[2] https://www.octoparse.com/
[3] https://www.scrapestorm.com/
[4] https://phantombuster.com/product

on Instagram has data such as the number of followers and the number of accounts followed. A portrait of an Instagram account is shown in Figure 1.

Besides, users who have an account can publish content in the form of images or videos where each content can have attributes such as caption, owner id, number of likes, number of comments, post address, time of publication, and image address various sizes. A portrait of an post from an account on Instagram is shown in Figure 2.

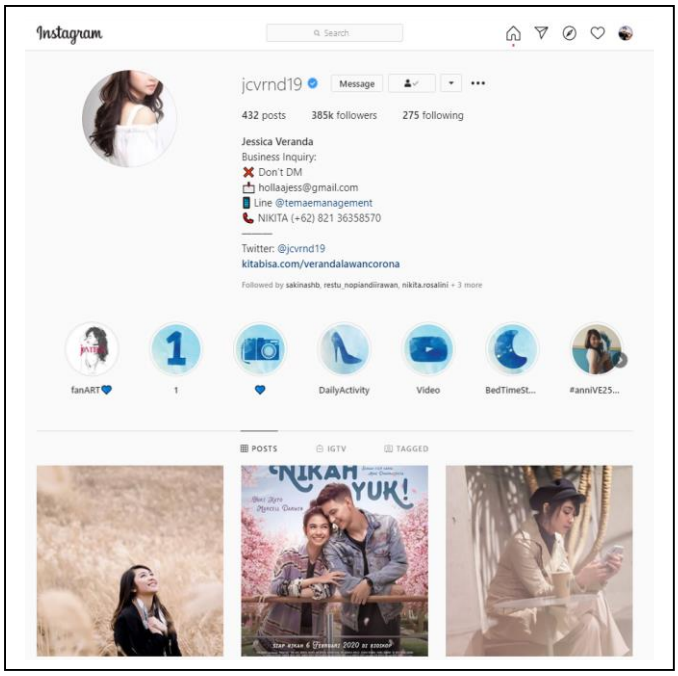

Figure 1. Example of an Instagram account

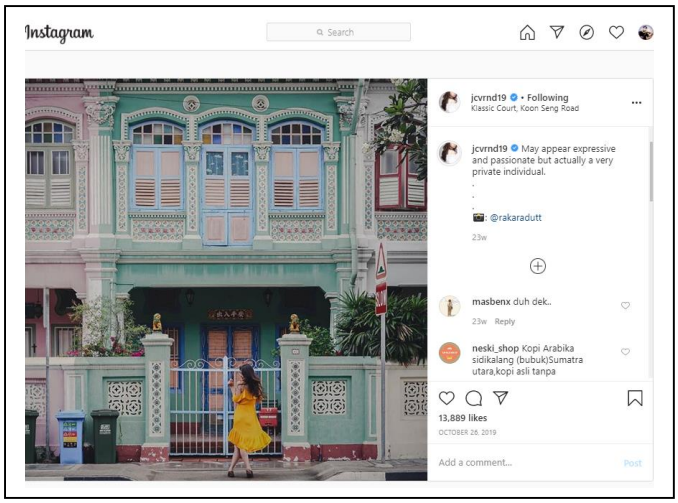

Figure 2. Example of an Instagram post

2.2 *Web Scraping Method Implementation*

Web scraping technique is a process for extracting data from the internet [17]. The extraction process is carried out to extract Instagram account data. The extraction process stages using this web scraping technique in this study are as follows [18]:

- In the analysis phase, the researcher studies the HTML and JSON structure of the Instagram website at this stage. This process aims to determine the data structure and elements to be downloaded from an Instagram account.
- The next process is to make the crawl done using a library called Beautiful Soup, which is found in the python programming language.
- The data extraction process on the Instagram web using web scraping techniques is carried out by sending a request to an Instagram web page address then extracting JSON (JavaScript Object Notation) data containing data from an Instagram account. The fourth step is to take Instagram account data in the form of captions, owner id, number of likes, and number of comments, postal address, post time, and image address. The data extraction process on the Instagram web with web scraping techniques is shown in Figure 3.

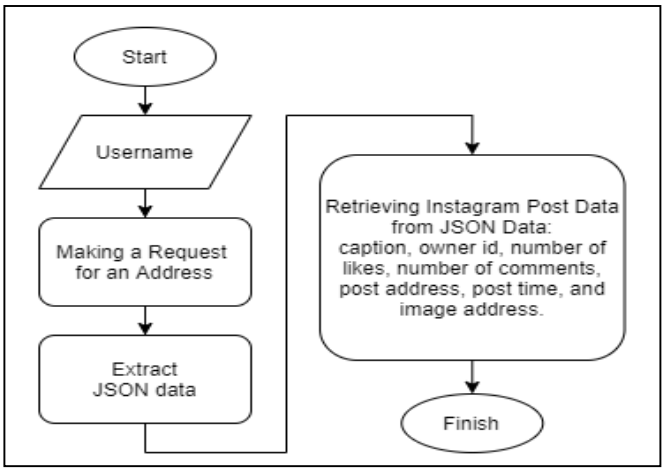

Figure 3. The extracting data process from the Instagram web

2.3 *Application's Architecture Design*

This application's architecture design consists of three main parts: Media Crawler, Data Repository, and Web User Interface. The architectural design of this application is shown in Figure 4.

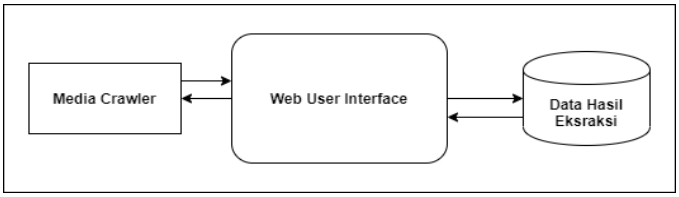

Figure 4. The extracting data process from the Instagram web

The functions of each of these sections in detail are as follows:

- Media Crawler. This section consists of a crawler program whose function is to extract data from the Instagram website pages. This section is implemented by utilizing web scraping techniques using libraries in the Python programming language called Beautiful Soup.
- Data Repository. This section serves as a storage area for extracted data. In this application architecture, the approach used in managing data in the repository is the NoSQL concept. This concept is a data management concept that is less schema, so it is suitable for the structure and characteristics of Instagram account data in unstructured text.
- Web User Interface. This section serves as a media interface between users and this application. Users can directly input several parameters on a. Through this module, users can also manage the list of Instagram account data that has been extracted directly. This web-based application utilizes a library for web application development called Flask.

2.4 Analysis of Application User Interaction

In this study, user interaction analysis is described in the form of a use case diagram. The use case diagram of the Instagram account data download application in this study is shown in Figure 5. Users in this application can download Instagram account data to the data collection and then manage it, such as deleting and exporting data collections in CSV, Excel, or JSON.

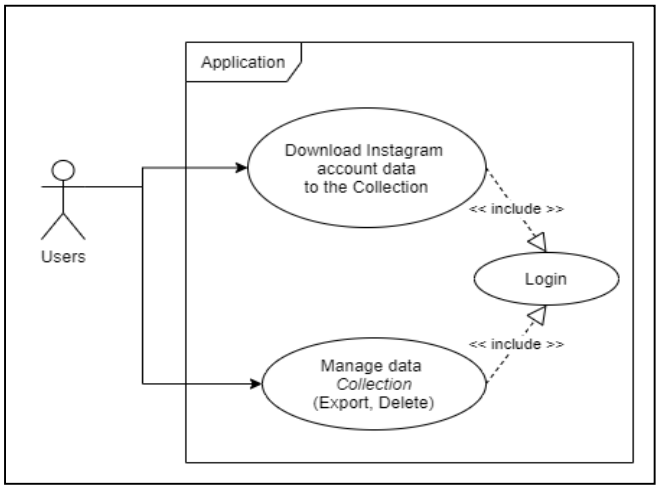

Figure 5. Application user interaction

2.5 Evaluation of Data Download Application

In this study, application evaluation was carried out by two testing methods, namely the black-box testing method and the download analysis test method. Testing with the black-box way is carried out by performing a functional test of each application page [19]. The download analysis test method is carried out to measure the extent to which the web scraping method is applied to applications that have been built in downloading Instagram account data. The Instagram account that will be used as an experiment in testing the analysis of the results of downloading account data consists of 15 accounts with a varying number of publications ranging from accounts that have 144 to 10420 post.

3 RESULT AND DISCUSSION

This chapter explains the results and discussion related to testing the Instagram data download application. The results of testing applications with black-box are shown in Table 1. The next evaluation is to analyze or measure the success of implementing the web scraping method to download Instagram account data. The trial phase aims to see the extent of this application in downloading Instagram account data. The trial for downloading Instagram account data on this application was conducted on Saturday, July 25, 2020, with the results shown in Table 2.

Table 1 Features Evaluation Result

No	Features	Evaluation Result
1	Download without criteria for the number and time of publication	Success
2	Download with number criteria	Success
3	Download with post time criteria	Success
4	Download with criteria for the post number and post time	Success
5	Export data to Excel	Success
6	Export data to CSV	Success
7	Export data to JSON	Success
8	Delete data	Success

Based on the black-box test results in Table 1, it can be seen that all the features in this application can run successfully. These features include data collection management, which functions to delete and export data collections in the form of CSV, Excel, or JSON. Based on the results of testing the analysis of the downloaded data in Table 2, it can be seen that the application of the web scraping method to download Instagram account data can download a maximum of 2412 publications of account data.

Table 2 Evaluation Results of Downloads

No	Username	Number of Post	Number of Downloaded Post
1	adripriadana	144	144
2	anna_yamada_	248	248
3	jcvrnd19	469	469
4	natgeoasia	864	864

5	freediverlife	932	932
6	explore_selayar	1072	1072
7	explorebima	1351	1351
8	explorelombok	1928	1928
9	explorebandung	2524	2412
10	exploreindonesia	3119	2412
11	bbcearth	4871	2412
12	natgeoadventure	5,779	2412
13	wonderful_places	6813	2412
14	tribunjogja	8262	2412
15	bbcnews	10420	2412

This Instagram account data downloader application consists of two main interface displays. The first interface is the Instagram account data downloader interface. In this interface, there is a form that the user uses to enter the Instagram account username from which the data will be downloaded. On this page, there is also a form to input the number of publications to be downloaded and enter the start date when the publication data will be downloaded where both inputs are optional. The downloaded account data will automatically be stored in the MongoDB database in the form of a collection. The interface for the Instagram account data downloader is shown in Figure 6.

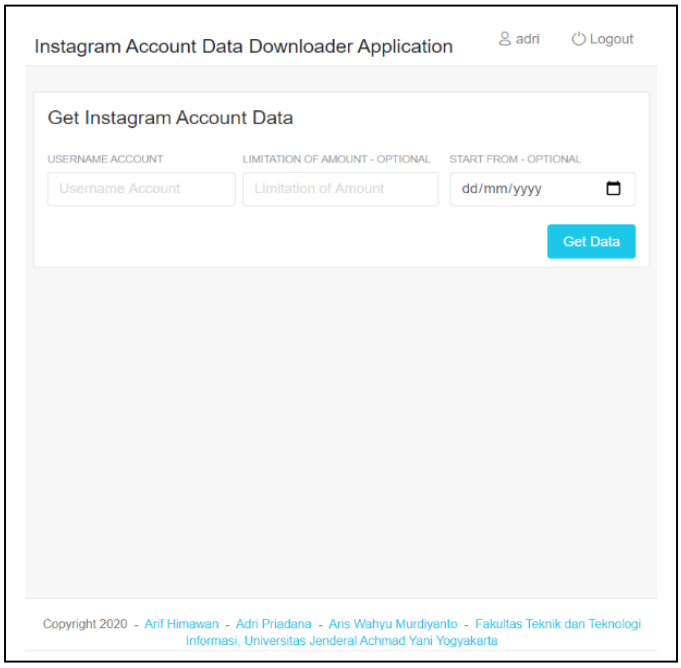

Figure 6. Instagram Account Data Downloader Interface

The second interface is the data collection management interface. In this interface, users can manage data from several Instagram accounts that have been downloaded previously. In this interface, users can delete and export data collections in A Comma Separated Values (CSV), Microsoft Excel, or JavaScript Object Notation (JSON). The data collection management interface is shown in Figure 7. Compared with the four parties offer Instagram account data download services, which was mentioned in the introduction, this application has advantages and disadvantages. Table 3 shows a comparison of the packaging of each application. Based on the comparison in Table 3, it can be seen that the four applications have usage limitations even though, in this case, the cost chosen is the cost of the cheapest package.

This application can be used for researchers, especially researchers who have data mining research, to download Instagram account data so that the next stage can be carried out, such as the data analysis stage. Not only for research purposes in data mining but also for research in big data, which is used to analyze social media [20] such as Instagram. Several other parties can also use this application for various purposes without having to pay (free).

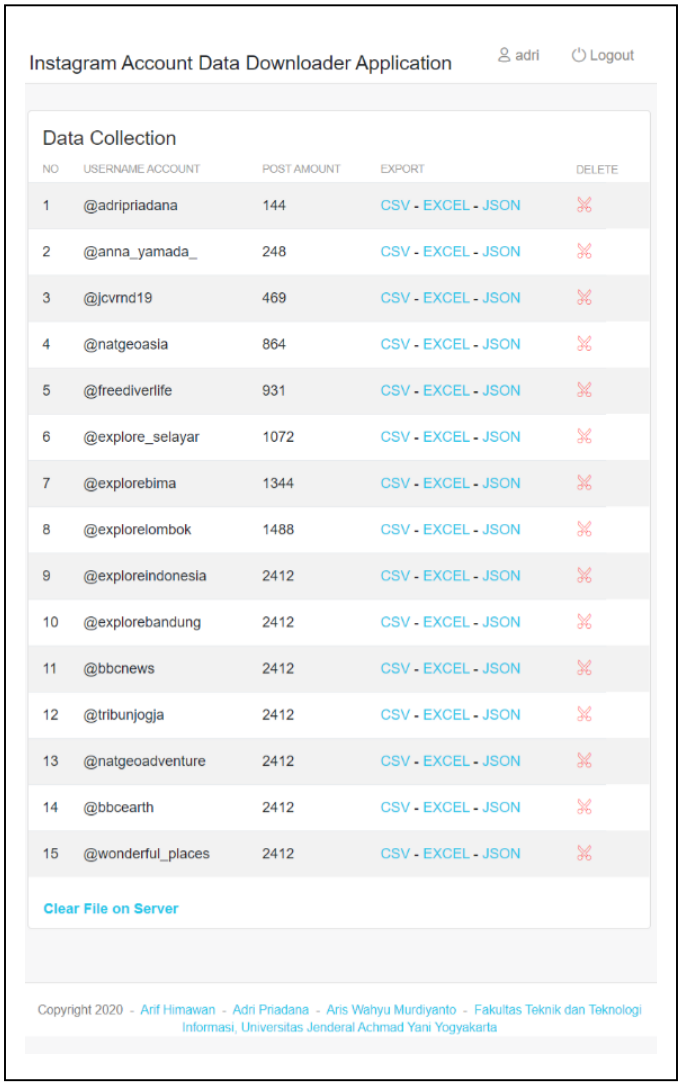

Figure 7. Data Collection Management Interface

Table 3 Comparison of the Advantages and Disadvantages

No	Application Name	Packaging
1	ScrapyGram	$50 / pack up to 1000 records to scrape.
2	Octoparse	$75 / month for unlimited pages per crawl.
3	ScrapeStorm	$49,99 / month for unlimited pages per task to scrape.
4	PhantomBuster	$30 / month (1 hour per day) for 5 task to scrape.
5	This application	Free for up to 2412 records to scrape with no time limit.

4 CONCLUSION

This research has successfully implemented the web scraping method to build an Instagram account data downloader application. In this study, application testing was carried out on 15 Instagram accounts with various publications, namely between 100 and 11000. Based on the results of the analysis of downloaded data, the web scraping implementation successfully to download Instagram account data was able to download a maximum of 2412 accounts post data. In this application, users can download Instagram account data to a data collection and then manage it, such as deleting and exporting data collections in CSV, Excel, or JSON.

ACKNOWLEDGMENT

The author would like to thank the Faculty of Engineering and Information Technology, Universitas Jenderal Achmad Yani Yogyakarta, for support in internal research funding assistance in the 2020 Applied scheme.

REFERENCES

[1] Y. Hu, L. Manikonda, and S. Kambhampati, "What we instagram: A first analysis of instagram photo content and user types." The AAAI Press, pp. 595–598, 2014.

[2] J. Constine, "Instagram hits 1 billion monthly users, up from 800M in September | TechCrunch," 2018. [Online]. Available: https://techcrunch.com/2018/06/20/instagram-1-billion-users/. [Accessed: 19-Dec-2019].

[3] S. L. B. Ginting, "Algoritma Apriori untuk Menampilkan Korelasi Nilai Akademik dengan Kelulusan Mahasiswa: Data Mining," *Komputika J. Sist. Komput.*, vol. 6, no. 2, pp. 59–65, Jun. 2019, doi: 10.34010/komputika.v6i2.1706.

[4] K. Latifah, "ANALISIS DAN PENERAPAN ALGORITHMA C45 DALAM DATA MINING UNTUK MENUNJANG STRATEGI PROMOSI PRODI INFORMATIKA UPGRIS," *J. Tek. Inform.*, vol. 11, no. 2, pp. 109–120, Nov. 2018, doi: 10.15408/jti.v11i2.6706.

[5] H. Sa'dyah, W. Sarinastiti, and R. R. Ramadhan, "Rancang Bangun Mesin Crawler di Instagram dan Pinterest untuk Kebutuhan Data pada Riset Visual," *MIND J.*, vol. 4, no. 1, pp. 24–37, Sep. 2019, doi: 10.26760/mindjournal.v4i1.24-37.

[6] B. Yadranjiaghdam, S. Yasrobi, and N. Tabrizi, "Developing a Real-Time Data Analytics Framework for Twitter Streaming Data," in *Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017*, 2017, pp. 329–336, doi: 10.1109/BigDataCongress.2017.49.

[7] R. Fauziah, I. A. Ratnamulyani, and A. A. Kusumadinata, "EFEKTIFITAS PROMOSI DESTINASI WISATA REKREASI GUNUNG PANCAR MELALUI POSTINGAN INSTAGRAM MEDIA SOSIAL," *J. Komun.*, vol. 4, no. 1, Jul. 2018, doi: 10.30997/jk.v4i1.1210.

[8] A. A. Arman and A. P. Sidik, "Measurement of Engagement Rate in Instagram (Case Study: Instagram Indonesian Government Ministry and Institutions)," in *Proceeding - 2019 International Conference on ICT for Smart Society: Innovation and Transformation Toward Smart Region, ICISS 2019*, 2019, doi: 10.1109/ICISS48059.2019.8969826.

[9] M. I. Akrianto, A. D. Hartanto, and A. Priadana, "The Best Parameters to Select Instagram Account for Endorsement using Web Scraping," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2019, pp. 40–45, doi: 10.1109/ICITISEE48480.2019.9004038.

[10] I. D. Utama and T. Inayati, "Brand Post Analysis and Categorization in Automobile's Instagram Accounts," in *Proceedings of 2019 International Conference on Information Management and Technology, ICIMTech 2019*, 2019, pp. 12–17, doi: 10.1109/ICIMTech.2019.8843753.

[11] W. Kurniawan, F. Ramadhan, and H. Ardiansyah, "The Application of Intersection in the Set Theory for Instagram Hashtags," *IJID (International J. Informatics Dev.*, vol. 8, no. 2, p. 88, Mar. 2020, doi: 10.14421/ijid.2019.08207.

[12] M. Azmi, Amiruddin Khairul Huda, and Arief Setyanto, "PEMANFAATAN DATA INSTAGRAM UNTUK MENGETAHUI REPUTASI TEMPAT WISATA DI LOMBOK," *Tek. Teknol. Inf. dan Multimed.*, vol. 1, no. 1, pp. 39–46, May 2020, doi: 10.46764/teknimedia.v1i1.13.

[13] A. Priadana and A. W. Murdiyanto, "Analisis Waktu Terbaik untuk Menerbitkan Konten di Instagram untuk Menjangkau Audiens," *J. Penelit. Pers dan Komun. Pembang.*, vol. 24, no. 1, pp. 59–70, Jun. 2020, doi: 10.46426/jp2kp.v24i1.118.

[14] A. Priadana and M. Habibi, "Face detection using haar cascades to filter selfie face image on instagram," in *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, 2019, pp. 6–9, doi: 10.1109/ICAIIT.2019.8834526.

[15] H. Ting, W. W. P. Ming, E. C. de Run, and S. L. Y. Choo, "Beliefs about the use of Instagram: an exploratory study," in *International Journal of Business and Innovation , 2 (2)*, 2015, pp. 15–31.

[16] A. Alsaeed, O. Alotaibi, N. Alotaibi, and M. Almutairy, "Automating

Instagram Activities and Analysis: A Survey of Existing Tools," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11578 LNCS, pp. 267–277, doi: 10.1007/978-3-030-21902-4_19.

[17] R. C. Pereira and T. Vanitha, "Web Scraping of Social Networks," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 3, no. 7, pp. 237–240, 2015.

[18] Fatmasari, Y. N. Kunang, and S. D. Purnamasari, "Web Scraping Techniques to Collect Weather Data in South Sumatera," in *Proceedings of 2018 International Conference on Electrical Engineering and Computer Science, ICECOS 2018*, 2019, doi: 10.1109/ICECOS.2018.8605202.

[19] M. Huda, S. Wiyono, M. F. Hidayatullah, and S. Bahri, "Studi Kasus: Sistem Informasi dan Pelayanan Administrasi Kependudukan," *Komputika J. Sist. Komput.*, vol. 9, no. 1, pp. 59–65, Apr. 2020, doi: 10.34010/komputika.v9i1.2518.

[20] N. Buslim, "Pengembangan Algoritma Unsupervised Learning Technique Pada Big Data Analysis di Media Sosial sebagai media promosi Online Bagi Masyarakat," *J. Tek. Inform.*, vol. 12, no. 1, pp. 79–96, Jun. 2019, doi: 10.15408/jti.v12i1.11342.